拡張アンサンブル法を用いた self-avoiding walk の数の推定 (次世代計算科学の基盤技術とその展開)

AUTHOR(S):

白井, 伸宙

CITATION:

白井, 伸宙. 拡張アンサンブル法を用いた self-avoiding walk の数の推定 (次世代計算科学の基盤技術とその展開). 数理解析研究所講究録 2013, 1848: 108-123

ISSUE DATE:

2013-08

URL:

http://hdl.handle.net/2433/195084

数理解析研究所講究録
第1848巻 2013年 108-123

拡張アンサンブル法を用いたself-avoiding walkの数の推定

白井 伸宙
大阪大学大学院 理学研究科, 大阪大学 サイバーメディアセンター.

概要

本稿ではモンテカルロ法を用いて巨大集合の要素数を推定する方法について解説する。数える対象は正方格子・立方格子上の有限ステップself-avoiding walkであり、厳密な数え上げの限界は正方格子で71ステップ、立方格子で36ステップである。どちらの数もアボガドロ数を上回る大きな数であるが、厳密に数える事を諦めれば、もっと大きな数を数える事ができる。我々は拡張アンサンブル法を用いたself-avoiding walkの数の推定法を2種類開発し、正方格子で256ステップ、数にして$6.2(4)\times 10^{108}$ (括弧内は最後の桁の標準誤差)、立方格子で200ステップ、数にして$3.6(1)\times 10^{134}$という、非常に大きな数まで数えることができた。統計誤差付きの推定では今まで達成されていない大きさである。これらの結果とともに、数を数えるための基本テクニックと拡張アンサンブル法を用いたレアイベントサンプリングについて詳しく解説する。

1 はじめに: 数えきれないくらい多いものを数える

数は有限だが、手で数えるには絶望的に多く、数え方を間違えれば計算機でも生きているうちには答えを返してくれないくらい大きな「数」について考えてみる。それは、塵劫記[1]にある大数の名を借りれば、那由他(10^{60})や不可思議(10^{64})、無量大数(10^{68})といった数である。現代を代表する大数で言えばGoogol[*1] (10^{100})であろう。これらの近隣の宇宙の原子数より多い何かを「数えたい」と思ったら、どのような方法がありうるだろうか。

「数える」という単純なことでも、数える対象や数えたい数の大きさ、必要な精度によって様々な方針が存在する。本稿では特に、数学的に定義された巨大集合の要素数を数えるのだが、まずは簡単のため、モノとして存在する何かを数える「数え方」について触れてみたい。松良氏の「動物の個体数調査法」[2]に習って、数え方を全数調査法と間接調査法の二つに分けてみよう。

全数調査法では、名前の通り「全ての数」をそのまま扱う。動物の個体数を例に取れば、ある指定した地域の動物をシラミ潰しに全部数えるという方法だ。銀行員がお金を数えるのも、昔ながらの選挙の開票も全数調査である。見逃しや二重数えなどの数え間違いがない限り厳密な数がわかる一方、扱う数が大きくなると、現実的な時間で数え上げることが難しくなってくる。ここで登場するのが、もう一つの方法、間接調査法である。

[*1] Google社の社名の元となった単語らしい。

間接調査法では数える対象すべてを扱うことを諦め、比較的少数の要素を抽出し、間接的に個体数を把握しようとする。そして間接調査法の一種である標本調査法では、母集団から標本（サンプル）を抽出し、サンプルが持つ統計的な性質から元の母集団の数を推定する。

本研究ではモンテカルロ法を用いた標本調査により母集団の数を推定する二つの手法を開発した。本稿のタイトルにある**拡張アンサンブル法**はモンテカルロ法の一種であり、珍しいが重要度の高い状態を効率よく抽出する（レアイベントサンプリング）ことができる。開発した二つの手法では self-avoiding walk (SAW) という格子上で数学的に定義された集合を扱うのであるが、どちらの手法もこのレアイベントサンプリングをうまく活かすことで大きな数を推定することができるようになっている。

2 Self-avoiding walk とその周辺

2.1 Self-avoiding walk (SAW) とは？

数え方の詳しい説明に入る前に、本研究で数える対象だと述べた SAW について説明する。

SAW とは、原点から伸びる格子の上の一筆書きのパスのうち「同じ点を 2 度通らない」という条件を満たすものである [3]。パスの両端を閉じれば self-avoiding polygon (SAP) と呼ばれる多角形になる。

記号を使って定義しておこう。d 次元格子上の点の集まりを $\omega = (\omega(0), \omega(1), \cdots, \omega(N))$, $\omega(i) \in \mathbb{Z}^d$ $(i = 0, 1, 2, \cdots, N)$ と表した時、$\omega(0) = (0,0)$ 及び $|\omega(i+1) - \omega(i)| = 1$ $(i = 0, 1, \cdots, N-1)$ の条件を満たすものとして N ステップのランダムウォークが定義される。このランダムウォークに $\omega(i) \neq \omega(j)$ という条件をすべての i, j $(i \neq j)$ の組について課せば N ステップ SAW となり、加えて $\omega(N) = \omega(0)$ を課すと N ステップ SAP となる。

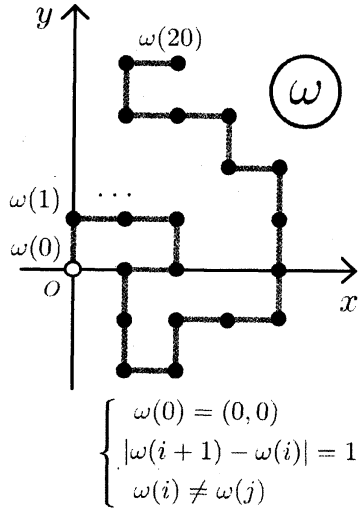

図 1 Self-avoiding walk を表すための記号と条件

方眼紙に書かれた線が SAW もしくは SAP であるかを目で見て判定するのは容易であるし、空間的な制約条件を満たしながらこれらの線を描くこともまた容易い[*2]。しかし、任意の長さの SAW/SAP について数学的に厳密に証明された定理は、ランダムウォークのそれに比べてはるかに少ない[*3]。

例えば、正方格子の原点から伸びる N ステップのランダムウォークの数は、簡単な考察から 4^N であるとわかるのに対し、SAW については、その一般形を知るための糸口さえつ

[*2] 我々の研究室では、このような SAP の性質を利用して囲碁に似た陣取りゲーム「Polygo」を作成した。ルール等の詳しい情報は以下の URI 参照のこと。http://www.cp.cmc.osaka-u.ac.jp/~shirai/jp/polygo.html

[*3] SAW/SAP ともに、5 次元以上で N 無限大の漸近挙動がランダムウォークと同じになることが知られており、数学的・物理的に興味が持たれるのは 4 次元以下、本稿で扱うのは 2 次元（正方格子）と 3 次元（立法格子）である。

かめていない。得られているのは、洗練された手法と十分な計算資源により達成された、たった 71 ステップまでの厳密数え上げの結果である [4]。その数は優にアボガドロ数を超え、「地道に数えるには多すぎる」と思わせる巨大な数字となっている。

2.2 鎖状高分子モデルとしての self-avoiding walk—格子タンパク質モデルを例に

2.2.1 SAW の応用

SAW は鎖状高分子の最も簡単なモデルとして興味が持たれていたため、厳密に解析することが困難でも、化学者や物理学者による近似計算や数値計算が古くから行われてきた。今日では、彼らによって得られた様々な予測—例えば長ステップ極限での物理量の漸近挙動に関する予測—が多く存在している。

SAW を鎖状高分子のモデルと考えた場合、同じ点を 2 度通らないという制約は分子鎖が有限の太さを持つという排除体積効果の一つの表現として考える。さらに分子内・分子間の相互作用などを加えることで、鎖状高分子の溶媒置換に対する振る舞いや熱力学的な振る舞いに関する解析に用いられることが多い。本節では鎖状高分子の中でも特にタンパク質を例に SAW の応用について説明する。

2.2.2 タンパク質の特殊性

タンパク質は生体「高分子」といえど、人工的に作られた一般の高分子とはまったく異なる性質を持っている。熱力学的な性質も異なれば、ミクロな状態にも違いがある。この違いがうまれた理由はタンパク質が生物と進化の過程を共にしてきた分子であるからであり、小さいながら、生物が生きるのに必要な機能の一部が備わっているからである。それらの機能と密接に関わっているのがタンパク質の**構造**である。

タンパク質とは 20 種類のアミノ酸を DNA に記された配列に従ってつないで作られる鎖状高分子である。タンパク質の多くは三次元的な構造を持ち[*4]、その構造はアミノ酸配列が同じであれば同じである。これらの構造はタンパク質が合成されてから折れ畳みという過程を経て構成され、折れ畳んだタンパク質は熱力学的に安定である。通常の鎖状高分子が、熱力学的に安定な同一構造をミクロに取ることなどほとんどありえず、このようなタンパク質の特殊性を表現した SAW のモデルに、格子タンパク質モデルがある。

格子タンパク質モデルの研究は 1970 年代に郷らの研究 [6, 7, 8] によって始まった。現在郷モデルと呼ばれているそのモデルは、タンパク質の折れ畳みや熱力学的な振る舞いを記述する理論的な基礎付けを与えている。ここでは 1975 年の論文で扱われた格子モデルを取り上げ、追試で行った計算機実験の結果とともに格子タンパク質及び郷モデルについて説明する。

格子タンパク質モデルでは、SAW をタンパク質分子だと考え、各ステップをアミノ酸

[*4] 「構造を持たないタンパク質」は長年きちんと解析されて来なかったが、1990 年の終わり頃から注目を浴び始め、**天然変性タンパク質**と呼ばれるようになった。現在に至るまで、精力的な研究が行われている。

残基[*5]だと思い、アミノ酸間の相互作用を仮定する。N ステップの SAW は残基数 M が $N+1$（数えていなかった 0 ステップ目分だけ増える）の格子タンパク質に対応していることに注意されたい。1975 年の論文では A, B, C の 3 種類のモデルが用意されており、同じ長さの SAW（$N=48$, $M=49$）にそれぞれが異なった分子内相互作用が仮定されている。モデル A, B は「タンパク質らしさ」を備えたモデルであり、モデル C は比較のために用意された通常の高分子の振る舞いを表現しているモデルである。

2.2.3 郷モデル

モデル A は現在郷モデルと呼ばれている相互作用を持っており、モデル B はモデル A の相互作用にランダムな相互作用が付け加えられている。まず郷モデルについて説明しよう。

郷モデルは「理想タンパク質」とも言うべきモデルであり、タンパク質が生体内で持つ構造（ネイティブ構造）がエネルギー的に安定になるように分子内相互作用が構成される。簡単に言うと、基底状態がネイティブ状態で、ネイティブ構造に近い状態ほどエネルギーが低く、遠いほど高い。

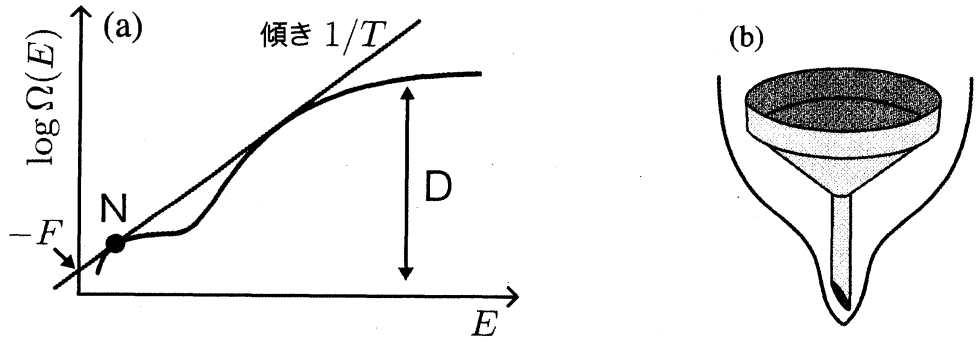

図 2　ファネル型のエネルギー地形について。(a) 郷モデルの状態密度の概念図。横軸にエネルギー E、縦軸に状態密度の対数 $\log \Omega(E)$ を取った時、N で示されたネイティブ状態と D で示された構造を持たない状態の間にへこみが存在し、高温から温度を下げていくとある温度を境に主要な状態が D から N へと遷移する。(b) (a) のグラフを立てて回転させると漏斗のような形になるため、「ファネル型のエネルギー地形」と呼ばれる。

郷モデルが持つ状態空間を端的に表しているのが図 2 (a) である。横軸にエネルギー E を、縦軸に状態密度[*6]の対数を取った時、エネルギーが高い状態（D）と低い状態（N）の間に「へこみ」が存在しており、高温から温度を下げていくと、ある温度で D から N へピョンと状態が遷移する。この状態遷移は**二状態転移**と呼ばれ、実際のタンパク質の熱測定で見られる現象を定性的に説明している[*7]。

[*5] タンパク質はアミノ酸がペプチド結合でつながってできる分子だが、タンパク質のうちの「もと一つのアミノ酸に属していた原子団」をアミノ酸残基と呼ぶ。

[*6] 状態密度とは単位幅のエネルギー範囲にどれくらいの状態が存在しているのかを表した物理量。例外は多くあるが、一般に低エネルギー状態より高エネルギー状態のほうが状態密度が高い。ここでは、この二つの間がどうつながれているかの話をする。

[*7] この二状態転移は、統計力学の言葉を使うと「一次相転移的な状態遷移」と言われる。一次相転移はマクロな系が示す熱力学的な転移のことであり、有限の長さのタンパク質を扱っているうちは相転移は起きない。

図 2 (a) を左に 90 度回転し、E 軸方向にぐるりと回すと、(b) に示したような漏斗形になる。「左に 90 度」や「ぐるりと回す」操作は実は謎であるが、とにかく出来上がった (b) の図のうち、x, y 軸を状態空間、z 軸をエネルギー E と思うことにすると、**ファネル型のエネルギー地形**と呼ばれるものになる。「ファネル」の語感が良いためか、タンパク質の理論的な話をする際には必ず顔を出す単語になっている[*8]。

郷モデルのこころは「タンパク質の折れ畳みや熱力学的な性質を記述する上でまず注目すべきなのはネイティブ構造周りの揺らぎである」という鋭い洞察に基づく仮説にあり、同様の仮定に基づいて作られたタンパク質の連続空間モデルはタンパク質の折れ畳みに関する実験をとても良く説明し、この仮説の有用性を示している [9]。

郷モデルではある特定のネイティブ構造を参照しつつ、そのネイティブ構造において距離が近いアミノ酸残基同士に引力相互作用を仮定する。これらのアミノ酸のペアをネイティブコンタクトペアと呼ぶ。郷モデルのハミルトニアンは一本の SAW (ω) に対して以下のように定義される。

$$\mathcal{H}(\omega) = - \sum_{(i,j)\in\{\text{ ネイティブコンタクトペア }\}} \varepsilon\, \delta(|\omega(i) - \omega(j)|, 1). \tag{1}$$

ここで、$\delta(x, y)$ はクロネッカーのデルタを表し、

$$\delta(x, y) = \begin{cases} 1 & (x = y) \\ 0 & (\text{otherwise}), \end{cases} \tag{2}$$

である。具体例をモデル A で紹介しよう。

モデル A は図 4 をネイティブ状態に、図 3 の黒い四角で指定された残基のペアをネイティブコンタクトペアに持つモデルである。図 3 で指定されたペアは、図 4 で表されている構造内で隣り合わせになっている残基のペアから、残基の番号が隣り合わせのペアを除いたものである。全部で 36 のペアがあるので、基底状態のエネルギーは $E = -36\varepsilon$ である。モデル B では図 3 の黒い四角に加えて、灰色のペアも加えた相互作用を持っている。灰色のペアはランダムに選ばれた相互作用ペアであり、「理想タンパク質」を表していたモデル A から少し現実のタンパク質に近づけることを意図して作られている。基底状態は A と同じ図 4 の構造であり、エネルギーは $E = -36\varepsilon$ である。

最後のモデル C は interacting self-avoiding walk (ISAW) と呼ばれる SAW で、分子鎖で隣り合わせの残基 (もしくはモノマー) 以外のすべての残基間に引力相互作用が存在する。このモデルではモデル A, B のように特定の構造と関連付けられていないので、7×7 の箱に収まるコンパクトな構造を取れば、そのどれもが基底状態となる。コンパクトな状態での相互作用する残基ペアは A, B と同じく 36 個であるので、基底状態のエネルギーは $E = -36\varepsilon$ である。

第 5 章ではいくらでも長いタンパク質のモデルを作れるモデルを扱い、長さが無限大の極限での振る舞いが一次相転移を示す。

[*8] 郷らによる "consistency principle" や Wolynes らの "minimum frustration" は同様の内容を含んでいるが、使用頻度は比較的低い。

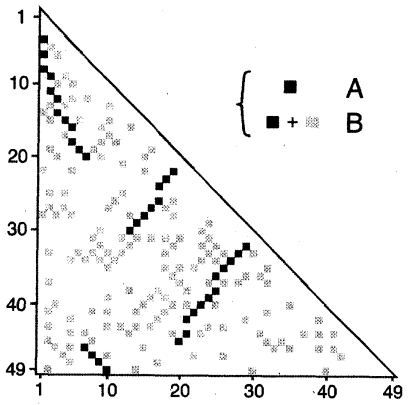

図 3　モデル A（黒）とモデル B（黒と灰色）のネイティブコンタクトペア。縦軸、横軸ともにアミノ酸残基の番号を示しており、黒・灰色で塗られた四角がネイティブコンタクトペア。ほぼ同じ図が 1978 年の郷・武富の論文 [7] に載っている。

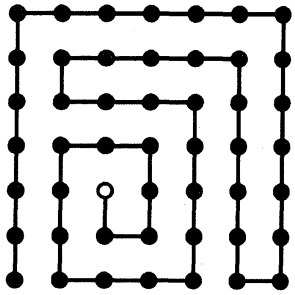

図 4　モデル A の基底状態（$E = -36\varepsilon$）。自明な対称性を除いてこの状態が唯一の基底状態であり、全てのネイティブコンタクトが形成されている。郷モデルではこのような状態を天然状態と呼ぶ。ほぼ同じ図が 1978 年の郷・武富の論文 [7] に載っている。

2.2.4　郷モデルの熱力学的性質

では、これらのモデルのシミュレーションから何が言えるのだろうか。ここでは拡張アンサンブル法を用いて求めた比熱のグラフ（図 5）を示している。

前述のとおりモデル A のエネルギー地形はファネル型になっている。比熱のグラフを見ると $T = 0.78$ に鋭いピークを持っているので、その前後で二状態転移を示すことがわかる。この温度をまたぐように温度を下げていくと、構造がなく揺らいだ状態から折り畳まれた状態へと遷移する。モデル B ではこのピークが少しなまるが、依然二状態転移的な振る舞いをしている。モデル C ではだらだらとしたグラフへと変化し、低温側の状態も多くの構造が縮退していて構造が定まらず、タンパク質らしさはない。

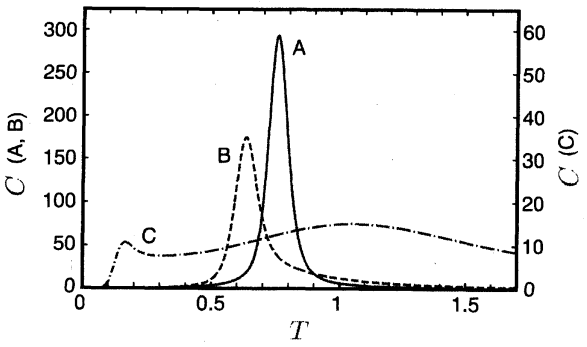

図 5　A, B, C のそれぞれのモデルに対する比熱のグラフ。A, B は左の軸、C は右の軸で比熱の大きさを表している。($\varepsilon = 1$)

3　数を見積もるための方策

第 1 章で少し触れたが、本研究では標本調査法を用いた 2 種類の数の推定法を開発した。本章では、それぞれの方法について詳しい説明を行う。

3.1 マーキング法と郷モデル

再び、松良氏の「動物の個体数調査法」を覗いてみる。そこには標本調査法の一種であるマーキング法（標識再捕獲法とも呼ばれる）が解説されている。例として上げているのは、池の中のフナが何匹いるかを数えるという問題である。

まず初めに、池から 100 匹のフナを無作為に抽出する。そして、そのフナの尾びれを少し切り取って目印を付け、再び池に放つ。放ったフナと他のフナが十分混ざったと思える頃に、もう一度フナのサンプルを抽出する。例えば 80 匹抽出した中に、先ほど目印をつけたフナが 20 匹混ざっていたとする。我々はここから、先の無作為抽出によって得られた 100 匹のフナが全体の四分の一（$20 \div 80$）を占めていたと予想し、池のフナの数が 400 匹であると推定する。

この方法の肝となっているのは、サンプリングを 2 回行い、1 回目のサンプリング時に比較可能な数（100 匹）を自ら用意するところである。全数を直接扱わずサンプルの統計に頼る手法の性質上、得られたサンプル内の個数比から相対的な量を推定すること—例えば 20cm より大きいものと小さいものの存在比—はできるが、母集団の個数を一回のサンプリングで得るのは難しい。しかし、ひと度自らの手で「スパイ」を紛れ込ませてしまえば、その「スパイ」が何人もしくは何匹いるのかについて自信を持って答えることができる。

同じ手を、SAW を数える問題に応用してみよう。ここで再び登場するのが郷モデルである。郷モデルではコンパクトなネイティブ構造を一つ仮定すれば、自明な基底状態とファネル型のエネルギー地形が現れるのであった。この自明な基底状態をスパイに仕立て上げるのが今回の作戦である。

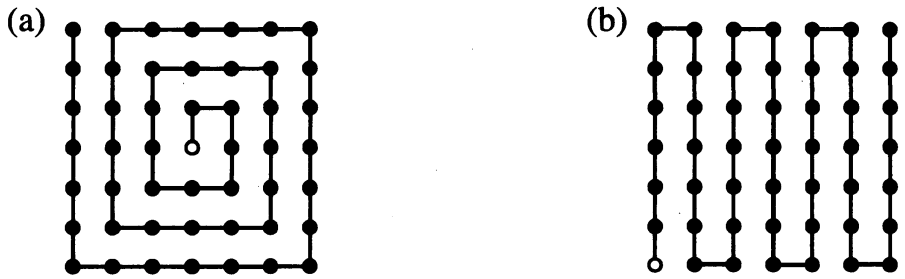

図 6　郷モデルの基底状態の例（$N = 48,\ M = 49$）。(a) ぐるぐる巻きの状態（「ロール構造」と呼ぶ) (b) ジクザグの状態（「ベータ構造」と呼ぶ）。白抜きの丸が原点。どちらの構造も任意の N に対して同様な構造を構成しやすい形になっている。

図 6 にあるようなネイティブ構造を仮定した時、基底状態は回転・反転の対称性により 8 つ存在し、厳密にわかる数として母集団の中に潜んでいる。これがフナの例でいうところの一回目のサンプリングに相当する。その後、SAW からのサンプリングを行い、この 8 つをうまく抽出できれば SAW の数を導き出すことができるのだが、そうは問屋が卸さない。SAW が長くなるに連れて、埋め込んだ 8 つの状態の相対量が指数関数的に少なくなっていくのである。

では、ステップ数が長くなるごとに指数関数的に多いサンプルを得なくてはならないかというと、実はまだ抜け道は残っている。鍵となるのは郷モデルのファネル型エネルギー地形である。現状において、何も、8つの基底状態がSAW全体の中にぽつんと浮かんでいるわけではなく、その基底状態に至るように徐々に数が絞られていく構造のエネルギー地形が存在している。これを使わない手はない。

第4章で説明するマルコフ連鎖モンテカルロ法では、SAWを部分的に変化させる変形規則を導入することで状態遷移列を作り出す。この状態遷移列を追っていくと、定義されている変形を使って何回でたどり着くことができるかを考えることで状態空間内で近い状態と遠い状態がおおまかに区別されるようになる。SAWを局所にしか変形しない規則を考えている場合、近い状態は同士は近いエネルギーを持っていることが期待できる。

この性質を利用すると、変形による状態遷移を繰り返し「あ、今エネルギーが下がったから、基底状態に近づいた気がする！」といったような感触を、エネルギーが変化するたびに感じることができるようになる。これは、広大な状態空間の中に大雑把な道標を散りばめられたことを意味する。この道標を使えば、「エネルギーの低い方へ」という方針に基づいてエネルギーを段階的に下げていくことで、基底状態を探し出すことができる。**レアイベントサンプリング**を可能にするための大事な仕掛けが、ここにある。レアイベントサンプリングの詳しい説明は第4章で行うとして、もう一方の数え方について説明しておこう。

3.2 モンテカルロ積分とDomb-Joyceモデル

もう一つの数の推定法で用いる数え方の方針は、円周率の計算を行う際のモンテカルロ積分に似ている。モンテカルロ積分では $-1 \leq x \leq 1,\ -1 \leq y \leq 1$ の正方形の中にランダムに点を打ち、単位円に入った点の数の割合から円周率を求めるのであるが、ここには絶対的な基準として、「面積が4の正方形」が用意されており、この基準を使ってまだ知らない円の面積にアプローチしている。前節では母集団の「一部」に厳密な数を用意したのだが、本節では母集団の「全体」に絶対的な基準を用意している。

SAWを数える問題でも同じ方針が使える。使うのはズバリ、**ランダムウォーク**である。正方格子の場合「面積4の正方形」に対応するのは「N ステップのランダムウォークの総数 4^N」であり、立方格子ならば総数は 6^N となる。ここで考えたいのは、SAWとランダムウォークの「つなげ方」である。

郷モデルを用いた数の推定法と同様、ランダムウォークの中に埋まっているSAWの割合はステップ数増加に従って指数関数的に落ちいてくので、レアイベントサンプリングが必要となる。そして、レアイベントサンプリングを用いるには状態空間中で迷子にならない道標としてのエネルギー構造が必要となるのであった。本研究では1972年にSAWの解析のために考案されたDomb-Joyceモデル [10] を用いてエネルギー構造を取り入れた。Domb-Joyceモデルはランダムウォークのモデルであり、格子上の同じ点を複数回通るたびにエネルギーが高くなる。このSAWの「重なり度」を V と表すと、Domb-Joyceモデルのハミルトニアンは

$$\mathcal{H}(\omega) = \sum_{i<j} J\ \delta(\omega(i), \omega(j)) = JV\ (J > 0), \tag{3}$$

のように表される。高温極限でランダムウォーク、低温極限でSAWとなるモデルになっており、基底状態では重なりが一つもない状態、すなわちSAWそのものとなる。

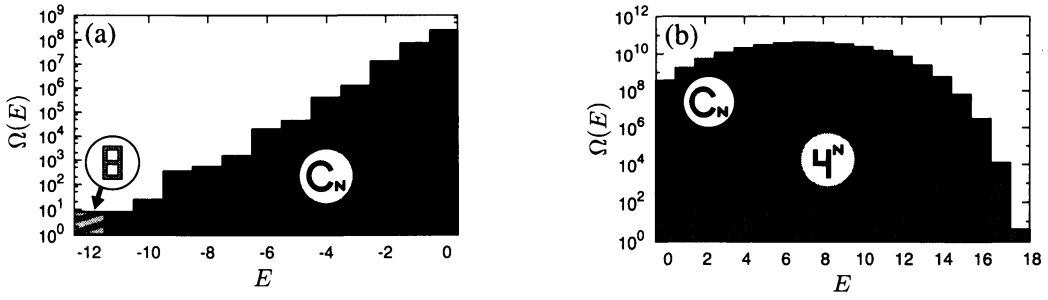

図7 (a) 郷モデル（ロール構造）、(b) Domb-Joyceモデルの状態密度。郷モデルの状態密度はファネル型、Domb-Joyceモデルの状態密度は釣り鐘型をしている。図中の c_N はSAWの数を意味する。

Domb-Joyceモデルのエネルギー地形は郷モデルのようにファネル型にはなっていない。郷モデルとDomb-Joyceモデルの状態密度を図7に示しているが、状態密度 $\Omega(E)$（対数表示）は最も状態数が大きい所から基底状態に向かい、へこんでいるのではなく、膨らんでいて、釣り鐘型をしている。この違いはすなわち、エネルギーが低いレアな状態と、エネルギーが高く桁違いに多い状態の間のつなぐパイプの太さの違いを表しており、「パイプの太いほうが良いのではないか」という直観が働く。この直観が正しいことを、熱力学量的な違いも交えて第5章で説明する。

4 SAWのシミュレーション

4.1 マルコフ連鎖モンテカルロ法

第3章で説明した推定法では、それぞれ

1. SAWの中からネイティブ構造を含んだサンプルを得る
2. ランダムウォークの中からSAWを含んだサンプルを得る

ことが必要だった。そして「ネイティブ構造」と「SAW」はそれぞれ「レア」な状態であるため、レアイベントサンプリングに頼る必要があると述べた。ここでは、レアイベントサンプリングを説明する準備として、そもそもSAWやランダムウォークからのサンプリングを行うにはどうすれば良いかについて解説する。まずはランダムサンプリングから考えてみよう。

ランダムウォークの場合はいたって簡単である。正方格子上であれば1～4の乱数を N 個用意し、それぞれの数字に上下左右をあてがって数字の列のとおりに原点から線を引けば、N ステップのランダムウォークからのランダムなサンプルが一つ得られる。例えば各サンプルについて重なり度 V を計算し、ヒストグラムを作ることで状態密度 $\Omega(V)$ を推定することができる。

SAW についてはどうだろうか。最も単純な方法として、「ランダムウォークをたくさん作って、SAW の条件を満たしたものだけ拾い上げる」という方法がある。しかし、ステップ数 N が大きくなるに従い、ランダムウォークの中に占める SAW の割合は指数関数的に小さくなるので、大きな N を持つ SAW のサンプルを一つ得るだけで一苦労である。この研究の目的は大きな数を推定することにあり、このままではいけない。

ここで登場するのが、本節のタイトルとなっている**マルコフ連鎖モンテカルロ法**である。SAW のマルコフ連鎖モンテカルロでは、長さ一定の鎖を図 8 に示すような部分的な変形、場合によっては全長にわたる大きな変形を重ねて時刻とともに変化させる。

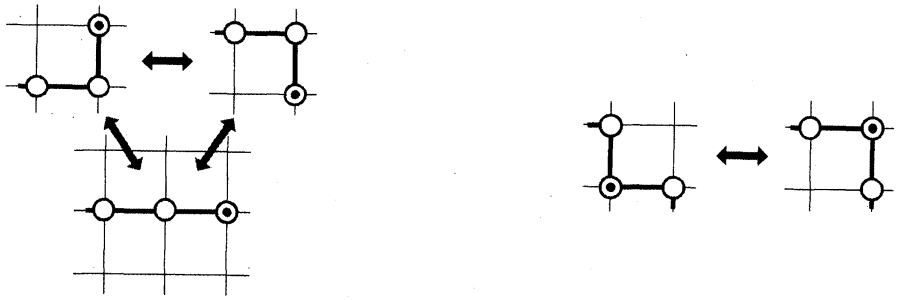

図 8　(a) SAW の先端・末端の変形規則、(b) SAW の中間ステップの変形規則。中黒の付いている丸を局所的に動かす。

この遷移列で現れる状態間の相関が時間に対して指数関数的に減少している時、減衰率の数倍の時間間隔で SAW を抽出すれば、サンプル同士の時間相関は十分小さくなると期待できる。エルゴード性[*9]が満たされていれば、最終的に集められたサンプルの集合は SAW からのランダムなサンプルとみなせる。相関がほんの少し残っているかもしれない状況は少し気持ち悪いかもしれないが、SAW を一つ作るたびに大量のランダムウォークを破棄しなければならない状況は緩和される。そして、状態空間を局所的に探索するこの方法について考えることで、**重み付きサンプリング**への道が開ける。

4.2　重み付きサンプリング

第 3 章で少し触れたが、マルコフ連鎖モンテカルロ法で局所的な状態遷移を重ねていけば、状態空間に近い/遠いの距離のようなものが現れ、近い状態が近いエネルギーを持っていることを期待すると、好きなエネルギー領域へにじみ寄っていくことができる。

重み付きサンプリングでは、状態が持つエネルギーごとに重要度を切り分け、サンプリングの割合を変化させる。例えば統計力学のカノニカルアンサンブルではボルツマン因子 $(\exp(-E/k_BT))$ によって重み付けされたアンサンブルを扱うため、重み関数 $W(E)$ を $\exp(-E/k_BT)$ に設定し、高温ではどのエネルギー状態もまんべんなく、低温では低エネルギー状態を重点的にサンプリングして平衡状態のアンサンブルを構成する。さらに、どの

[*9] ある変形規則を用いた状態遷移列について、全ての状態へ至る道が用意されている性質のこと。マルコフ連鎖モンテカルロ法がきちんとしたアンサンブルを作り出すための条件の一つ。

ような重み付けを行ったのかを知っているので、得られたヒストグラムを重み関数の逆数 $(1/W(E) = \exp(E/k_BT))$ をかけて元に戻してやると、状態密度 $\Omega(E)$ が推定できる。では、低温のカノニカルアンサンブルを作り出すことで、基底状態にあるレアなサンプル達をうまく拾いあげることができるだろうか？

「レアな状態を見つけ出せるか」という問いならば、答えは YES である。例えば郷モデルをある程度低温にしておけば、基底状態を見つけ出すことは比較的容易い。Domb-Joyce モデルに関しても同じである[*10]。しかし、「レアな状態の、レアでないものに対するレア度」をうまく見積もれるかというと、そうではない。特に郷モデルでは、図 5 A のグラフでいうピーク温度前後で、平衡状態を特徴づける典型的な状態がほどけた状態（高温）から折れ畳んだ状態（低温）へと変化してしまうため、低温のシミュレーションをしていたら、ほぼ基底状態しか出ない。「レア」なものしか出ないなら、その「レア度」はわからないのである。[*11]

ではピーク温度あたりでカノニカルアンサンブルを作れば良いのかというと、これもまた難儀である。ファネル型エネルギー地形のほどけた状態と折れ畳んだ状態の間はくびれていて遷移途中の状態の出現確率が低いので、思ったようにこの二つの状態の間を行き来してくれない。

実は、レアなものをサンプルしつつレア度も同時に測れる、もっと良い方法がある。大事なのは一端、「温度一定のカノニカル分布」といった具体的な物理的状況を想定するのをやめることである。重み付けの仕方は、何もボルツマン因子に限ることはない。途中どんな重みを使おうと、最後にかけた重みを元に戻してやれば状態密度を求める事ができる。必要があればこの状態密度にボルツマン因子をかけて、好きな温度のカノニカルアンサンブルを作ることもできる[*12]。もっと自由な重み付けを考えてみよう。この「目の前にある物理をありのままに扱わない」というやり方が「拡張」アンサンブル法の考え方に直接通じている。

4.3 拡張アンサンブル法

4.3.1 マルチカノニカル法

自由な重み付けといっても、当然ながらそのやり方は無数にある。その中で広いエネルギー範囲の状態密度を精度よく求められる方法、ラフに言うなら、「レアなもののレア度」をうまく評価できる方法はどのようなものだろうか。

この問いに対して、良い指針を与えたのが Berg らのマルチカノニカル法である [11, 12]。マルチカノニカル法における大事な発想その一が、「重み関数 $W(E)$ をそれぞれの問題にオーダーメイドで作ってしまおう」である。はじめからどんな問題にも適用できる重み関数を考えるのは大変だけれど、問題に合わせた重み関数を、ある一定の計算コストを払ってこ

[*10] 多くのエネルギー極小状態が存在し、エネルギー地形が複雑に入り組んでいるモデルではこの二つの例のようにうまくはいかない。

[*11] 「生命が誕生した星がどれくらい珍しいか」や、「我々が存在する宇宙はどれくらい珍しいか」という問いを考えた時にも、同じ問題が生じる。我々は、地球以外の生命体について知らず、また今いる宇宙以外の宇宙は以外は知らないので、珍しい気がするのだけど、珍しさを定量的に評価できない。

[*12] この方法はリウェイティングと呼ばれる。

しらえてあげる方が、いくらか融通が利くようになる。

そして、大事な発想その二は、「重み関数 $W(E)$ を、状態密度の逆数 $\log\Omega(E)$ に（だいたい）比例するように用意しよう」である。本来未知である $\Omega(E)$ が現れるのに違和感を感じるかもしれないが、ニュートン法のような反復計算と同じく、種となる重み関数からスタートして、徐々に $\Omega^*(E)$ を練り上げていき、だいたい $\Omega(E)$ となった $\Omega^*(E)$ の逆数を重み関数 $W(E)$ として用いる[*13]。カノニカルアンサンブルを作る時の重み関数はボルツマン因子 $\exp(-E/k_BT)$ であったが、それを $1/\Omega^*(E)$ に置き換えるのだ。

状態密度の逆数に比例するように重みを設定するということは、つまりは「たくさんある状態の重みは小さく、少ない状態の重みは大きく」ことを意味する。「少ない状態の重みは大きく」などと言われると、「レアイベントサンプリング」のにおいがぷんぷんするではないか。

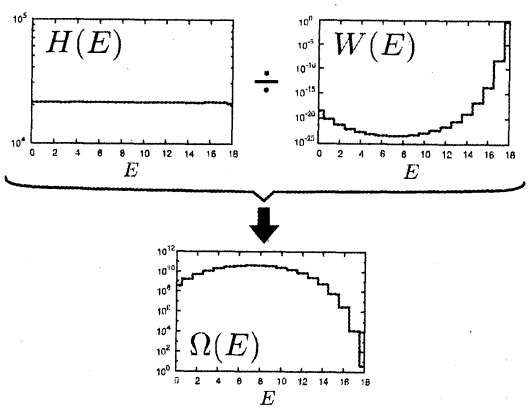

図 9　状態密度の逆数に比例した重み関数 $W(E)$ と平らなヒストグラム $H(E)$ を使うと状態密度 $\Omega(E)$ が精度よく推定できる。図に用いたのは $N = 19$ の Domb-Joyce モデルのマルチカノニカル法による計算。

重み関数の作り方についてここでは詳しく説明しないが、菊池氏と千見寺氏の文章 [19, 20, 21] に詳しい解説が載っているので、そちらを参照されたい。オーダーメイドの重み関数が状態密度の逆数にだいたい比例するように得られたとして、その重み関数に従ってサンプリングを行うと、何が得られるのだろうか。その特徴が最もわかりやすいのは、エネルギーに対して取ったヒストグラム $H(E)$ である。ランダムサンプリングでは $\Omega(E)$ に比例するように得られるヒストグラムに $1/\Omega^*(E)$ の重みがかかっているため、ヒストグラムは平ら（$H(E) \propto \Omega(E) \times 1/\Omega^*(E) \propto$ だいたい const.）になる（図 9 左上）。ヒストグラムを重み関数 $W(E)$（図 9 右上）で割ると、エネルギー軸の広い範囲で精度の良い状態密度 $\Omega(E)$ が求まる（図 9 下）。

この方法を用いると、低エネルギー状態も高エネルギー状態も、その間にあるどのエネルギー状態も同じ数だけサンプルが得られる。低温の郷モデルのサンプリングのように、エネルギーが低い状態しか出ないという状況は改善され、比熱のピーク温度で行うサンプリングのように、二状態転移の両側をうまく行き来できないような状況も、その間を繋ぐ細いパイプの重みを適度に高めているため、やはり改善される。

マルチカノニカル法により得られるマルチカノニカルアンサンブルは、実際の物理系で想定される温度一定の系とはかけ離れた系への拡張が行われている。この意味で、マルチカノニカル法は**拡張アンサンブル法**の一種であるといわれる。ここでは、本研究と深く関わって

[*13] オーダーメイドで状態密度の逆数にだいたい比例する重み関数を作成するにはいろいろやり方があって、Berg と Neuhaus の方法の他に Lee [13] の entropic sampling、Wang と Landau による Wang-Landau 法 [14] などが存在している。

いるもう一つの拡張アンサンブル法、multi-self overlap ensemble (MSOE) [17] について紹介しよう。

4.3.2 Multi-self-overlap ensemble (MSOE)

MSOE は格子タンパク質・格子ポリマーの系の統計力学的解析のために作られた方法で、SAW の条件を少し緩めて重なりを許し、Domb-Joyce モデルのような重なり度 V を導入する。本研究においても、郷モデルを用いて SAW の数を推定する際に MSOE を用いており、ここでは郷モデルを例に説明を行う。

MSOE ではマルチカノニカル法で扱っていた重み関数 $W(E)$ を更に重なり度 V 方向にも拡張して $W(E,V)$, $(V \leq V_{\mathrm{cut}})$ とする。ここで V_{cut} は重なり度 V のカットオフであり、扱う SAW の長さに合わせて適当な大きさに設定する。郷モデルは SAW で定義されているモデルであり、本来 $V=0$ の状態以外は存在しないのだが、状態空間を重なり方向に少し「拡張」し、$0<V\leq V_{\mathrm{cut}}$ の状態も許すようにしている。オーダーメイドの重み関数を作る方針は $W(E,V)\propto 1/\Omega(E,V)$ であり、この $W(E,V)$ を使うと 2 次元の平らなヒストグラム $H(E,V)$ が得られる。ヒストグラムが得られた後は、$H(E,V=0)/W(E,V=0)$ とすることで精度の良い状態密度 $\Omega(E,V=0)$ が得られる。

MSOE が目指す方針を一言で表すと「損して得取れ」である。状態空間を広げ、探索すべき空間を広げてしまう「損」の代わりに、平衡状態への緩和を速めたり、エルゴード性を満たせるようにしたりという「得」がある[*14]。MSOE で取るこの方針は、格子タンパク質のだいたいの問題に対してマルチカノニカル法をそのまま使うより得が多いようである。

これらの拡張アンサンブル法を用いたレアイベントサンプリングを用いると、レアなサンプルが拾えて、レア度もわかる。あとは第 3 章で示した方針の通りに数を推定するのみである。

5 計算結果

本研究では、1. 正方格子上の SAW の数を郷モデルと Domb-Joyce モデルの 2 種類の方法を用いて数え、2. 立方格子上の SAW を Domb-Joyce モデルによって数えた。(1. の結果については白井・菊池 [22] で示しているものとほぼ同じ。) 推定の結果を表 1, 2 に示している。

どれだけ大きな数を数えられたかというと、一番大きいもので立方格子上の SAW、20 ステップの 3.6×10^{134} である。「Googol (100^{100}) のアボガドロ数倍以上大きい数」などと表現することはできるが、もはや日常的な感覚ではよくわからないくらい大きな数になっている。

表 1 では三つの N について SAW の数を示しているが、$N=143$ が郷モデルを用いた推定法の限界[*15]であり、Domb-Joyce モデルの限界 $N=256$ に遠く及んでいない。この推定

[*14] では V_{cut} をどの程度大きく取ればエルゴード性を満たすのかと問われると、すぐには答えられない。ただ、$V\mathrm{cut}=N-2$ まで許すと、確実にエルゴード性を満たせる。

[*15] マルチカノニカル法で重み関数 $W(E)$ がうまく作れなくなる限界。

表 1　開発した手法を用いて推定したステップ数 N の 2 次元 SAW の数（表中の括弧は標準誤差を表す）

N	厳密解	Gō	Domb-Joyce
71	4.191×10^{30}	$4.3(2) \times 10^{30}$	$4.20(5) \times 10^{30}$
143		$1.4(4) \times 10^{61}$	$1.19(8) \times 10^{61}$
256			$6.2(4) \times 10^{108}$

表 2　Dom-Joyce モデルを用いて推定したステップ数 N の 3 次元 SAW の数（表中の括弧は標準誤差を表す）

N	厳密解	Domb-Joyce
36	2.941×10^{24}	$2.940(8) \times 10^{24}$
200		$3.6(1) \times 10^{134}$

法の性能差はどこからくるものなのだろうか？

3.2 節では郷モデルと Domb-Joyce モデルの状態密度の形を比べて大まかに分けた二つの状態集団をつなぐ「パイプの太さ」について触れ、Domb-Joyce モデルの方が性能がよさそうだと述べた。直観的な説明はこれで尽きているが、この直観が他の物理量でどのように見えるか、またステップ数を変化させたとき、直観で想定した状況からどう変化するかついて説明したい。

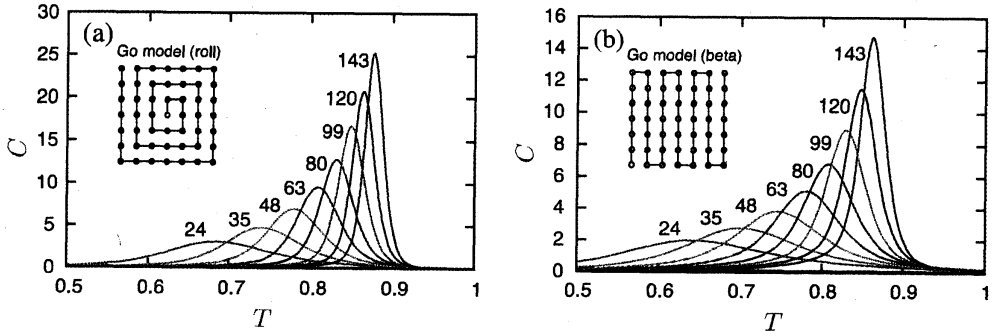

図 10　(a) ロール構造、(b) ベータ構造をネイティブ構造に取った時の郷モデルの比熱

状態密度が計算できたら、そのまま比熱の計算が可能となる。図 10、11 にはそれぞれ郷モデルと Domb-Joyce モデルについて、N をさまざま変えて書いた比熱のグラフを示している。ここからわかるのは、郷モデルではロール構造・ベータ構造ともに、$N \to \infty$ の極限で**一次相転移**を起こすであろうことと、Domb-Joyce モデルにはその転移が見られないことである。

この一次相転移とは、ある温度を境にカノニカルアンサンブルを構成する主要な状態集団が変化する転移であり、高温側からまたぐと、高エネルギーで状態数が膨大な状態集団（エントピックに安定）から低エネルギーで状態数が比較的少ない状態集団（エネルギー的に安定）への遷移が起こる。二状態転移と異なるのは、一次相転移は無限に大きい系での熱力学的な

転移を表していることである。つまり郷モデルのSAWを長くしていくと、二状態間の遷移はある温度で突然におきるようになり、二状態をつなぐ「パイプ」は相対的に細くなっていく。

マルチカノニカル法ではこのパイプをオーダーメイドの重み関数で太くすることを試みるのであるが、そもそもパイプの入り口をなかなか見つけられない場合、太くしようがなくなる。しかし、Domb-Joyceモデルにはこの転移がなく、パイプの入り口が十分広くなっている。つまり、郷モデルよりもパイプの入り口が見つけやすくなっている。そしてこの性質は長いステップにしても変わらない。パイプは太いまま残っている。以上の議論から、図7で説明した「ファネル型か、釣り鐘型か」という状態密度の違いは重み関数$W(E)$を作るコストを大きく変化させ、後者の方が簡単なのではないだろうか[*16]。

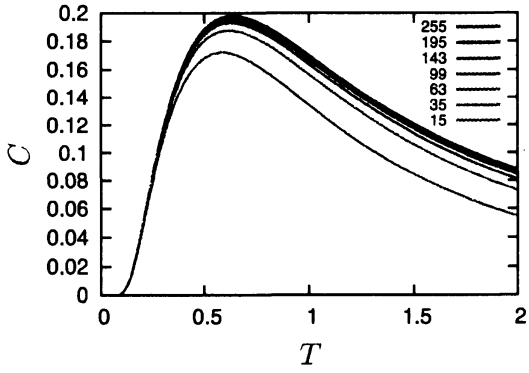

図11　Domb-Joyceモデルの比熱

以上の議論を積極的に解釈するなら、SAWを数える問題に限らず、任意のエネルギー構造を埋め込んで、レアイベントサンプリングを用いて数を勘定する問題では、釣り鐘型のエネルギー構造を埋め込んだほうが良いと言えそうだ。

6　最後に

「SAWを数える」というごく単純な研究の内容であるが、2種類のアプローチで推定法を構成し、レアイベントサンプリングの使いどころなども詳しく説明した。ある意味で二つの推定法の優劣ははっきりしているのであるが、同種の方法でSAW以外の数の推定を行う問題を考える際、手段は複数あった方がよく、またそのどちらもが使えるとき、真っ先に使うべきがどちらかについて述べることができた。さらに、より良いエネルギー構造を構成するための方針はどのようなものかを検討する良い材料になるのではないかと思う。

本原稿を読んでコメントを頂いた菊池誠先生、降旗大介先生、小渕智之さん、話す機会及び、講究録を書く機会を与えてくださった研究代表者の山本有作さん、研究提案者の谷口隆晴さんに感謝いたします。

参考文献

[1] 吉田光由,『塵劫記』(1627).
[2] 松良俊明,「動物の個体数調査法」, 京都教育大学理科教育研究年報 **8** (1978) 1.
[3] N. Madras and G. Slade, *The Self-Avoiding Walk* (Boston, MA: Birkhäuser) (1993).

*16 とはいうものの、互いに全く異なるモデルであり、推定法の方針も異なるので、はっきりしたことは言えない。はっきりと言えることは、「SAWを数える問題では、郷モデルを使った推定法よりもDomb-Joyceモデルを使った推定法の方が性能が良かった」ということだけである。

[4] I. Jensen, *J. Phys. A: Math. Gen.*, **37** (2004) 5503.
[5] R. D. Schram, G. T. Barkema, and R. H. Bisseling, *J. Stat. Mech.* (2011) P06019.
[6] H. Taketomi, Y. Ueda, and N. Gō, *Int. J. Pept. Protein Res.*, **7** (1975) 445;
[7] N. Gō and H. Taketomi. *Proc. Natl. Acad. Sci. USA.*, **75** (1978) 559.
[8] N. Gō, *Annu. Rev. Biophys. Bioeng.*, **12** (1983) 183.
[9] Y. Levy, S. S. Cho, J. N. Onuchic, and P. G. Wolynes, *Journal of Molecular Biology*, **346** (2005) 1121.
[10] C. Domb and G. S. Joyce, *J. Phys. C: Solid State Phys.*, **5** (1972) 956.
[11] B. A. Berg and T. Neuhaus, *Phys. Lett. B*, **267** (1991) 249.
[12] B. A. Berg and T. Neuhaus, *Phys. Rev. Lett.*, **68** (1992) 9.
[13] J. Lee, *Phys. Rev. Lett.*, **71** (1993) 211.
[14] F. Wang and D. P. Landau, *Phys. Rev. Lett.*, **86**,(2001) 2050.
[15] F. Wang and D. P. Landau, *Phys. Rev. E*, **64** (2001) 056101.
[16] G. Chikenji, M. Kikuchi, and Y. Iba, *Phys. Rev. Lett.*, **83**,(1999) 1886.
[17] Y. Iba, G. Chikenji, and M. Kikuchi, *J. Phys. Soc. Jpn.*, **67** (1998) 3327.
[18] 伊庭幸人 ほか, 『計算統計 II マルコフ連鎖モンテカルロ法とその周辺』, 岩波書店 (2005).
[19] 菊池誠, 「モンテカルロで行こう!、または、ダイスをころがせ」物性研究 **63** (1994) 199.
[20] 菊池誠, 「モンテカルロ法のアヴァンギャルド：あるもののシミュレーションからないもののシミュレーションへ」物性研究, **71** (1999) 608.
[21] 千見寺浄慈, 「計算統計力学的手法による格子タンパク質模型の研究」, 物性研究, **73** (2000) 1025.
[22] N. C. Shirai and M. Kikuchi, Multicanonical simulation of the Domb-Joyce model and the Go model: new enumeration methods for self-avoiding walks, arXiv:1212.2181 (2012).